4-Ч

УДК 004.62:510.22

О.Р. Чертов, Д.Ю. Тавров
Національний технічний університет України «Київський політехнічний інститут»,
03056, Україна, пр-т Перемоги, 37, м. Київ, chertov@i.ua

Забезпечення групової анонімності в мікрофайлі з нечіткими даними

O.R. Chertov, D.Y. Tavrov
National Technical University of Ukraine "Kyiv Polytechnic Institute"
03056, Ukraine, Prospect Pobedy,37, Kiev, chertov@i.ua

Providing Group Anonymity in the Microfile with Fuzzy Data

О.Р. Чертов, Д.Ю. Тавров
Национальный технический университет Украины
«Киевский политехнический институт», г. Киев
03056, Украина, пр-т Победы, 37, г. Киев, chertov@i.ua

Обеспечение групповой анонимности в микрофайле с нечеткими данными

У статті пропонується спосіб використання нечіткої інформації в даних мікрофайла перепису населення для забезпечення їхньої групової анонімності. Розглядається метод підрахунку міри належності кожного статистичного запису множині записів, розподіл яких потрібно анонімізувати. Підхід ілюструється за допомогою розв'язання задачі групової анонімності на базі реальних даних.
Ключові слова: мікрофайл, групова анонімність, система нечіткого виведення.

In the paper, the method for utilizing fuzzy information in census microfile data for providing their group anonymity is proposed. A novel method for calculating membership grade for each statistical record belonging to a group to be anonymized is discussed. The approach is being illustrated with solving the group anonymity task based on real-life data.
Key Words: microfile, group anonymity, fuzzy inference system.

В статье предлагается способ использования нечеткой информации в данных микрофайла переписи населения для обеспечения их групповой анонимности. Рассматривается метод подсчета меры принадлежности каждой статистической записи множеству записей, распределение которых нужно анонимизировать. Подход иллюстрируется с помощью решения задачи групповой анонимности на основе реальных данных.
Ключевые слова: микрофайл, групповая анонимность, система нечеткого вывода.

Вступ

Сучасні інформаційні технології дозволяють аналізувати великі обсяги даних, які раніше важко було навіть уявити. Особливу цінність становлять первинні дані, на базі яких дослідник може висувати певні гіпотези та перевіряти їхню справедливість. Проект IPUMS-International [1], у рамках якого на момент написання даної роботи зібрано близько 397 млн записів про респондентів 185 переписів населення в 62 країнах, є одним із найпоказовіших прикладів доступності первинних даних для аналізу. Проте публікація первинних даних очевидним чином підвищує ризик розкриття такої інформації, яка повинна бути захищена, наприклад, певної інформації про особу.

Під знеособленістю чи анонімністю даних розумітимемо неідентифіковність суб'єкта в рамках певної множини суб'єктів [2]. Захист інформації про особу перед-

бачає виконання *індивідуальної* анонімізації даних, алгоритми якої розробляються в рамках аналізу даних зі збереженням приватності (privacy-preserving data mining), статистичного контролю за розкриттям інформації (statistical disclosure control) та інших суміжних областей. Натомість захист внутрішніх закономірностей, особливостей розподілу даних передбачає виконання їхньої *групової* анонімізації. Відповідні методи передбачають захист властивостей даних, які неможливо визначити шляхом аналізу окремих елементів даних.

Існуючі методи анонімізації працюють із числовою інформацією, яка не завжди адекватно відображає суть оброблюваних даних. Наприклад, значення категоріальних атрибутів мають вибиратися з наперед визначеної множини кодів, які часто не мають математичного сенсу. Оскільки ключовими елементами людської свідомості є нечіткі множини, тобто класи об'єктів, у яких перехід від присутності до відсутності поступовий [3], то дані демографічних спостережень – часто нечіткі, лінгвістичні за своєю природою. Тому доцільною видається їх обробка з застосуванням апарату нечітких множин.

Метою даної роботи є вдосконалення методів забезпечення групової анонімності статистичних даних для можливості анонімізації лінгвістичних даних.

Огляд літератури

Методи забезпечення індивідуальної анонімності повинні досягати двох основних цілей [4]: вони повинні як гарантувати захищеність інформації на достатньому рівні, так і забезпечувати лише несуттєве зниження корисності самих даних. Методи з цими властивостями відомі як методи публікації даних із забезпеченням приватності [5]. Їх можна розділити на *пертурбативні* (із збуренням) та *непертурбативні* (без збурення). Анонімність при застосуванні методів першої групи досягається шляхом унесення в дані певного збурення. До них належать такі підходи, як рандомізація, різновиди агрегації (*k*-анонімізація, *l*-різнорідність та *t*-близькість), обмін даними, застосування сингулярно-спектрального розкладення, невід'ємної матричної факторизації, вейвлет-перетворень та перетворень Фур'є. Методи другої групи виконують анонімізацію даних без спотворення. До них належать *рекодування даних* (їхнє укрупнення) та *стримання даних* (вилучення даних із початкового мікрофайла).

Окремо виділяють методи, які використовують при інтерактивному забезпеченні приватності. При цьому виникають специфічні проблеми, такі як порушення анонімізації з допомогою зовнішніх джерел. Для подолання цих складнощів успішно використовують концепцію диференціальної приватності [6].

У літературі виділяють різні класи задач забезпечення групової анонімності даних. Кількісну задачу групової анонімності як забезпечення анонімності кількісного розподілу деякої групи респондентів певною множиною значень (наприклад, розподіл військових різними регіонами держави) було розв'язано в [7]. У рамках кількісної задачі неможливо розв'язати задачі приховання розподілу відношень кількостей респондентів деякої групи до загального числа респондентів. Такі задачі мають назву концентраційних задач групової анонімності [8]. Однією з них можна вважати концентраційно-різницеву задачу [9], яка передбачає приховання розподілу різниці між двома концентраційними розподілами (наприклад, різниця в розподілах юнаків призовного віку та дівчат того ж віку різними регіонами може вказати на місцезнаходження військової бази).

Найповніше питання забезпечення групової анонімності розкрито в [10]. Загальну методологію групової анонімізації представлено в [11]. На її основі можливі доопра-

цювання існуючих або розробка нових методів забезпечення групової анонімності, у тому числі з використанням методів нечіткої логіки.

Поняття нечіткої множини [12] дозволяє вирішувати задачі, де джерелом недостатньої точності є відсутність чітко вираженого критерію належності об’єкта множині, а не випадковий характер вхідних даних. Природним розвиненням цієї ідеї є поняття лінгвістичної змінної [13], використання якої дозволяє подолати так званий принцип несумісності, за яким висока точність розв’язку несумісна з високою складністю розв’язуваної задачі. Складність систем, що пов’язані з діяльністю людини, зокрема демографічних і соціальних, на вивчення яких спрямовано проведення перепису населення, унеможливлює отримання точного результату при застосуванні стандартних методів аналізу механістичних систем.

Лінгвістична змінна, значеннями якої є нечіткі множини, є засадничим поняттям нечіткої логіки [14], яку розглядають у двох сенсах. У вузькому розумінні це – логічна система для формалізації приблизного міркування, яке дозволяє апроксимувати логічні судження, якими послуговується людина; у широкому – це теорія нечітких множин. У даній роботі нечітка логіка розглядається у вузькому сенсі.

Засоби нечіткої логіки використовують при розробці експертних систем, нечітких контролерів, у задачах розпізнавання образів, у медицині, економіці та інженерії [15]. Методи нечіткої логіки пропонують для індивідуальної анонімізації даних [16]. Проте на сьогоднішній день не існує математичних методів забезпечення групової анонімності статистичних даних за допомогою приблизних міркувань, незважаючи на принципово нечітку природу вхідних даних.

Схема забезпечення нечіткої групової анонімності

Під *мікрофайлом* $\mathbf{M}$ розумітимемо дані про респондентів, зібрані в один файл. Такий файл складається з записів u_i, $i=\overline{1,\mu}$, кожен з яких містить значення атрибутів w_j, $j=\overline{1,\eta}$. Елементи записів мікрофайла позначатимемо z_{ij}. Далі вважатимемо, що мікрофайл деперсоніфікований, тобто серед його атрибутів немає *ідентифікаторів*, що однозначно визначають респондента.

Позначимо через $\mathbf{w}_i$ множину всіх значень w_i-го атрибута. Тоді під *сутнісною множиною* $\mathbf{V}=\{V_1, V_2, \ldots, V_{l_v}\}$ розумітимемо підмножину декартового добутку значень *сутнісних атрибутів* $\mathbf{w}_{v_1}\times\mathbf{w}_{v_2}\times\ldots\times\mathbf{w}_{v_r}$, $v_i\in\mathbb{Z}$ $\forall i=\overline{1, r}$. Кожен елемент $V_k\in\mathbf{V}$, $k=\overline{1, l_v}$, $l_v\le\left|\mathbf{w}_{v_1}\right|\cdot\left|\mathbf{w}_{v_2}\right|\cdot\ldots\cdot\left|\mathbf{w}_{v_r}\right|$, де $|\cdot|$ позначає кардинальність множини, називають *сутнісною комбінацією значень*, яка в свою чергу складається з *сутнісних значень*. За допомогою сутнісних множин можна описати підмножини респондентів, розподіл яких потрібно приховати.

Параметризуюча множина $\mathbf{P}=\{P_1, P_2, \ldots, P_{l_p}\}$ – це підмножина $\mathbf{w}_p$, $p\ne v_i$ $\forall i=\overline{1, r}$. Відповідний атрибут називатимемо *параметризуючим*, а елементи множини $P_k\in\mathbf{P}$, $k=\overline{1, l_p}$, $l_p\le\left|\mathbf{w}_p\right|$ – *параметризуючими значеннями*. Їх використовують для впорядкування даних мікрофайла.

Під *групою* $G(\mathbf{V},\mathbf{P})$ розумітимемо множину, що складається з сутнісних комбінацій значень $\mathbf{V}$ та параметризуючих значень $\mathbf{P}$, визначених для деякої конкретної задачі групової анонімності [11].

Для роботи з нечіткими даними узагальнимо поняття сутнісної комбінації. На базі кожного атрибута мікрофайла визначимо лінгвістичну змінну L_i, універсальною множиною для якої є значення цього атрибута, а назва збігається з назвою атрибута. Значення змінної належать її терм-множині $T\left(L_i\right)$. Лінгвістична змінна може бути в певному сенсі виродженою, тобто її терм-множина може складатися повністю або частково зі значень, які можна бієктивно відобразити на універсальну множину.

Узагальнена сутнісна множина $\tilde{\mathbf{V}} = \{\tilde{V}_1, \tilde{V}_2, \ldots, \tilde{V}_{l_{\tilde{v}}}\}$ – це підмножина декартового добутку значень лінгвістичних змінних, визначених на сутнісних атрибутах, $T\left(L_{v_1}\right)\times T\left(L_{v_2}\right)\times\ldots\times T\left(L_{v_{\tilde{r}}}\right)$, $v_i \in \mathbb{Z}$ $\forall i = \overline{1, \tilde{r}}$. Кожен елемент $\tilde{V}_k \in \tilde{\mathbf{V}}$, $k = \overline{1, l_{\tilde{v}}}$, $l_{\tilde{v}} \le \left|T\left(L_{v_1}\right)\right|\cdot\left|T\left(L_{v_2}\right)\right|\cdot\ldots\cdot\left|T\left(L_{v_{\tilde{r}}}\right)\right|$ називатимемо *узагальненою сутнісною комбінацією значень*.

Під *нечіткою групою* $\tilde{G}(\tilde{V}, \mathbf{P})$ визначимо множину, що складається з узагальнених сутнісних комбінацій значень $\tilde{\mathbf{V}}$, а також параметризуючих значень $\mathbf{P}$, визначених для деякої конкретної задачі групової анонімності.

Останнє визначення дозволяє віднести будь-якого респондента з мікрофайла u_i, $i = \overline{1, \mu}$ до певної нечіткої групи шляхом обчислення функції належності $\mu_{\tilde{G}}(u_i)$ даного респондента. Множину значень $\mu_{\tilde{G}}(u_i)$ $\forall i = \overline{1, \mu}$ позначатимемо

$$\tilde{\mathbf{M}}_{\tilde{G}} = =\left\{\mu_{\tilde{G}1}, \mu_{\tilde{G}2}, \ldots, \mu_{\tilde{G}q}\right\}.$$

Під *задачею забезпечення групової анонімності* розумітимемо модифікацію початкового набору даних для кожної (нечіткої) групи $\tilde{G}_i(\tilde{\mathbf{V}}_i, \mathbf{P})$, $i = \overline{1, k}$ таким чином, щоб вразливі (для розв'язання даної задачі) властивості даних було захищено. Початковий набір даних $\mathbf{M}$ має бути змінений окремо для кожної (нечіткої) групи, щоб захистити специфічні особливості кожної з них. Загальна схема забезпечення групової анонімності передбачає виконання наступних кроків:

1. Підготовка деперсоніфікованого мікрофайла $\mathbf{M}$.

2. Визначення груп $\tilde{G}_i(\tilde{\mathbf{V}}_i, \mathbf{P})$, $i = \overline{1, k}$, що представляють категорії респондентів, розподіл яких потрібно захистити.

3. Для кожного i від 1 до k:

а) вибір *цільового представлення* $\Omega\left(\mathbf{M}, \tilde{G}_i\right)$, яке визначає набір даних довільного вигляду, що представляє особливості даної групи в початковому мікрофайлі в спосіб, зручний для його модифікації;

б) виконання відображення даних за допомогою *цільового відображення* Υ: $\mathbf{M} \rightarrow \Omega_i(\mathbf{M}, \tilde{G}_i)$;

в) одержання *модифікованого цільового відображення* за допомогою *модифікуючого функціоналу* Ξ: $\Omega_i(\mathbf{M}, \tilde{G}_i) \rightarrow \Omega^*{}_i(\mathbf{M}, \tilde{G}_i)$ таким чином, щоб вразливі особливості набору даних стали прихованими;

г) отримання модифікованого мікрофайла шляхом застосування *оберненого цільового відображення* Υ^{-1}: $\Omega^*{}_i(\mathbf{M}, \tilde{G}_i) \rightarrow \mathbf{M}^*$.

4. Підготовка модифікованого мікрофайла $\mathbf{M}^*$ до публікації.

Процедура побудови цільової поверхні

Розглянемо процес врахування нечіткої інформації при побудові конкретного вигляду цільового представлення – цільової поверхні, яку можна розглядати як узагальнення поняття *цільового сигналу*, уперше введеного в [7]. Під цільовим сигналом розуміють одновимірний впорядкований масив $\theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$, що відображає деякі властивості (нечіткої) групи. Його можна представити як функціональну залежність значень сигналу від параметризуючих значень P_k та деякого сталого параметра $\tilde{\mathbf{V}}$, що визначає набір (узагальнених) сутнісних комбінацій значень для даної (нечіткої) групи: $\theta_k = \theta\left(P_k, \tilde{\mathbf{V}}\right)$. У даній роботі розглядатимемо цільовий сигнал як *кількості* респондентів із певними сутнісними та параметризуючими значеннями – *кількісний сигнал* $q = \{q_1, q_2, \ldots, q_m\}$. Кількісний розподіл може призвести до порушення приватності, бо екстремальні його значення є вразливою інформацією.

За наявності нечітких даних, цільового сигналу виявляється недостатньо для представлення всієї інформації про респондентів у мікрофайлі. У цьому випадку можна ввести узагальнююче поняття *цільової поверхні*, під якою розумітимемо функцію двох змінних та сталого параметра $\Theta_{kj} = \Theta\left(P_k, \mu_{\tilde{G}j}, \tilde{\mathbf{V}}\right)$, $P_k \in \mathbf{P}$, $\mu_{\tilde{G}j} \in \tilde{\mathbf{M}}_{\tilde{G}}$, яка встановлює залежність значення на цільовій поверхні від параметризуючого та узагальнених сутнісних значень певного респондента, з одного боку, та значення функції його належності даній нечіткій групі – з іншого.

Розглянемо процедуру побудови кількісної цільової поверхні. Для кожного респондента u_i, що належить нечіткій групі, тобто значення сутнісних атрибутів w_j для якого належать універсальним множинам відповідних лінгвістичних змінних L_j, потрібно обчислити його функцію належності цій групі $\mu_{\tilde{G}}\left(u_i\right)$. Найзручнішим, на нашу думку, способом є використання системи нечіткого виведення (fuzzy inference system, FIS). У даній роботі використовуватимемо одну з найпоширеніших на практиці таких систем – систему типу Мамдані [17]. Типова система цього класу має набір вхідних та вихідних лінгвістичних змінних, а в основі логічного виведення лежить система нечітких продукційних правил вигляду

$$\text{якщо } x_1 \text{ є } A_{1k_1}, \ldots, x_n \text{ є } A_{nk_n}, \text{ то } y \text{ є } B_1,$$

де $x_1, \ldots, x_n$ та y – вхідні та вихідна змінна відповідно, $A_{1k_1}, \ldots, A_{nk_n}$ та B_1 – функції належності, визначені на вхідних та вихідній змінній відповідно. Функціонування системи складається з наступних кроків [18]:

1. На вхід системи надходять чіткі значення, до яких для підрахунку ступеня істинності антецедентів кожного правила застосовують функції належності, визначені на вхідних змінних. Цей етап називають *фазифікацією*.

2. Обчислене значення істинності антецеденту застосовується до консеквенту відповідного правила, формуючи таким чином єдину нечітку множину, що відповідатиме вихідній змінній цього правила.

3. Нечіткі множини, отримані для правил виведення, об’єднуються певним чином для отримання єдиної нечіткої множини, що відповідає вихідній змінній.

4. Отримана нечітка множина замінюється єдиним чітким числом, яке найкраще її відображає. Цей етап називають *дефазифікацією*.

Таким чином, для побудови системи нечіткого виведення потрібно:

1. Вибрати вхідні та вихідну лінгвістичні змінні та визначити на них функції належності для кожного значення лінгвістичної змінної.

2. Збудувати систему нечітких продукцій.

3. Вибрати спосіб реалізації нечіткого об’єднання, перетину та імплікації для обчислення нечіткої множини для кожного правила, а також – спосіб об’єднання нечітких множин для кожного правила в єдину множину.

4. Вибрати алгоритм дефазифікації.

Поклавши як вхідні змінні визначені раніше лінгвістичні змінні L_j, а як єдину вихідну змінну – ступінь належності конкретного респондента нечіткій групі, а також визначивши правила нечіткого виведення, антецедентами яких є узагальнені сутнісні комбінації значень, можна збудувати FIS-систему для підрахунку ступеня належності кожного респондента нечіткій групі. Варто зазначити, що, окрім нечітких правил, які складають ядро системи нечіткого виведення, предметна область також може додатково накладати певні (чіткі) обмеження на ті чи інші комбінації вхідних значень, які потрібно враховувати окремо від FIS-системи для підвищення достовірності остаточних результатів.

Для побудови кількісної поверхні потрібно підсумувати кількості респондентів із конкретним параметризуючим значенням та ступенем належності нечіткій групі:

$$\Theta_{kj} = \Theta\left(P_k, \mu_{\tilde{G}j}, \tilde{\mathbf{V}}\right) = \left|\left\{u_i \,\middle|\, z_{iw_p} = P_k,\ \mu_{\tilde{G}}\left(u_i\right) = \mu_{\tilde{G}j}\right\}\right|. \qquad (1)$$

Отримані значення потрібно впорядкувати певним чином. Наприклад, параметризуючі значення можна впорядкувати з міркувань їхньої природи, а ступені належності – за зростанням або спаданням.

Цільову поверхню майже неможливо коректно аналізувати, оскільки з погляду забезпечення групової анонімності цінність становлять кількості респондентів не з конкретними значеннями ступеня належності, а зі значеннями, що належать певному інтервалу. Тому додатково потрібно розбити множину ступенів належності $\tilde{\mathbf{M}}_{\tilde{G}}$ на інтервали $\Delta_{\tilde{\mathbf{M}}s}$ і підрахувати кількість респондентів, ступені належності яких потрапляють у відповідні інтервали:

$$\Theta_{ks} = \Theta\left(P_k, \mu_{\tilde{G}j}, \tilde{\mathbf{V}}\right) = \left|\left\{u_i \,\middle|\, z_{iw_p} = P_k,\ \mu_{\tilde{G}}\left(u_i\right) \in \Delta_{\tilde{\mathbf{M}}s}\right\}\right|. \qquad (2)$$

Практичні результати

Для ілюстрації описаного підходу до побудови кількісної цільової поверхні розглянемо приклад розв’язання задачі групової анонімності на базі реальних даних. Для можливості порівняння результатів побудови цільової поверхні з отриманими раніше результатами побудови цільового сигналу, розв’яжемо задачу нечіткої групової анонімності на базі прикладу, детально описаного в [7], у якому задача групової анонімності ставилася наступним чином: замаскувати територіальний розподіл військових, що працюють у штаті Каліфорнія, США. Як дані для аналізу було взято п’ятивідсоткову вибірку даних перепису населення США 2000 року [19]. Також було вибрано 16 статистичних областей, визначених для цього штату, з номерами 06010, 06020, 06030, 06040, 06060, 06070, 06080, 06090, 06130, 06170, 06200, 06220, 06230, 06409, 06600, 06700.

У даній роботі задачу нечіткої групової анонімності сформулюємо наступним чином: замаскувати територіальний розподіл осіб, яких можна вважати військовими з деяким ступенем упевненості (який називатимемо «мірою військовості» респондента), що працюють у штаті Каліфорнія. Для аналізу візьмемо дані, описані вище.

Як вхідні змінні для системи нечіткого виведення було взято «Вік», «Стать» та «Рівень освіти». Усі три змінні відповідають однойменним атрибутам мікрофайла. На наш погляд, комбінації цих атрибутів дають можливість найкраще оцінити ступінь належності респондента класу військових осіб. При цьому для спрощення підрахунку кількісної поверхні вважатимемо, що військовими можуть вважатися тільки чоловіки віком від 17 до 89 років.

Для лінгвістичної змінної «Вік» було взято три значення – «молодий», «середнього віку» та «похилого віку» (рис. 1).

Для змінної «Стать» було взято два значення – «чоловік» та «жінка». Оскільки кожну людину (за невеликими виключеннями, які не стосуються поточного дослідження) можна однозначно назвати представником тієї чи іншої статі, то дана змінна фактично є прикладом виродженої лінгвістичної змінної.

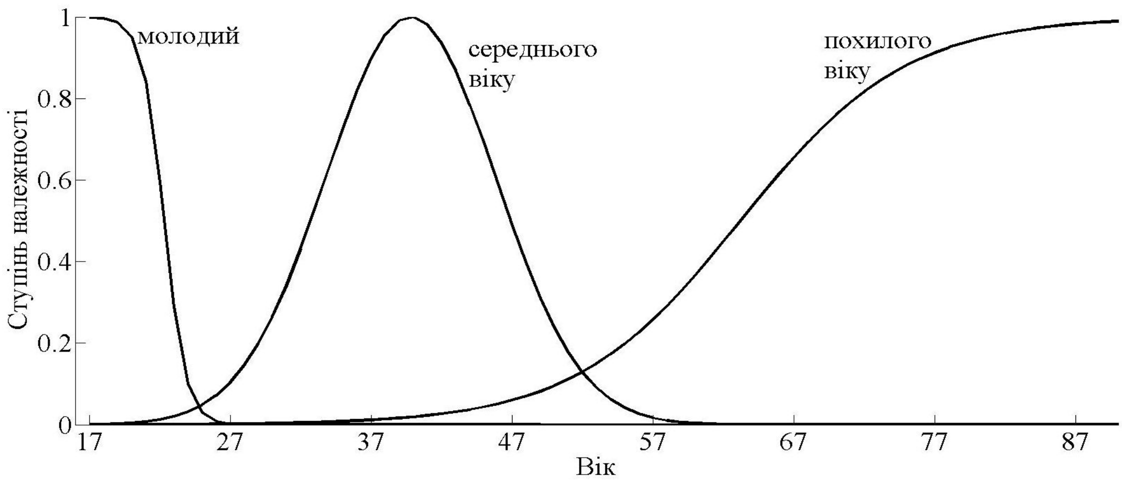


Рисунок 1 – Функції належності для лінгвістичної змінної «Вік»

Для змінної «Рівень освіти» було вирішено взяти два значення – «Низька освіта» та «Висока освіта», оскільки більша деталізація навряд чи суттєво впливає на можливість того чи іншого респондента вважатися військовим. Функції належності для даної змінної, з урахуванням категоріальної природи атрибута, представлено на рис. 2.

Як вихідну змінну було взято «Міру військовості» респондента, значеннями якої є «низька», «середня» та «висока» (рис. 3).

Наступним кроком побудови FIS-системи є визначення правил логічного виведення. Для даного прикладу були визначені наступні правила.

1. Якщо «Вік» = «молодий», «Стать» = «чоловік» і «Рівень освіти» = «низький», то «Міра військовості» = «висока».

2. Якщо «Вік» = «молодий», «Стать» = «чоловік» і «Рівень освіти» = «високий», то «Міра військовості» = «середня».

3. Якщо «Вік» = «середнього віку», «Стать» = «чоловік» і «Рівень освіти» = = «низький», то «Міра військовості» = «висока».

4. Якщо «Вік» = «середнього віку», «Стать» = «чоловік» і «Рівень освіти» = = «високий», то «Міра військовості» = «низька».

5. Якщо «Вік» = «похилого віку» і «Стать» = «чоловік», то «Міра військовості» = = «низька».

Як метод нечіткого об'єднання було взято функцію максимуму, перетину – мінімуму, імплікації – мінімуму, об'єднання нечітких множин – максимуму, дефазифікації – центру тяжіння.

Окрім побудованої FIS-системи, для підрахунку міри військовості того чи іншого респондента мікрофайла також було додатково враховано значення атрибута «Військова активність»: якщо значення дорівнює 1, то респондент вважається військовим із ступенем належності 1.

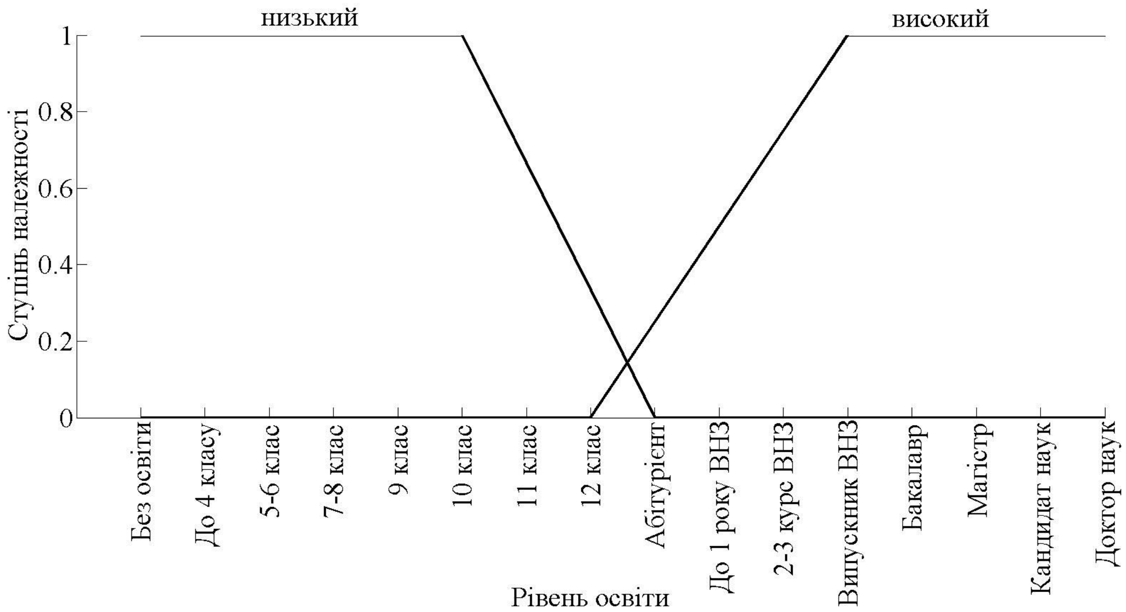


Рисунок 2 – Функції належності для лінгвістичної змінної «Рівень освіти»

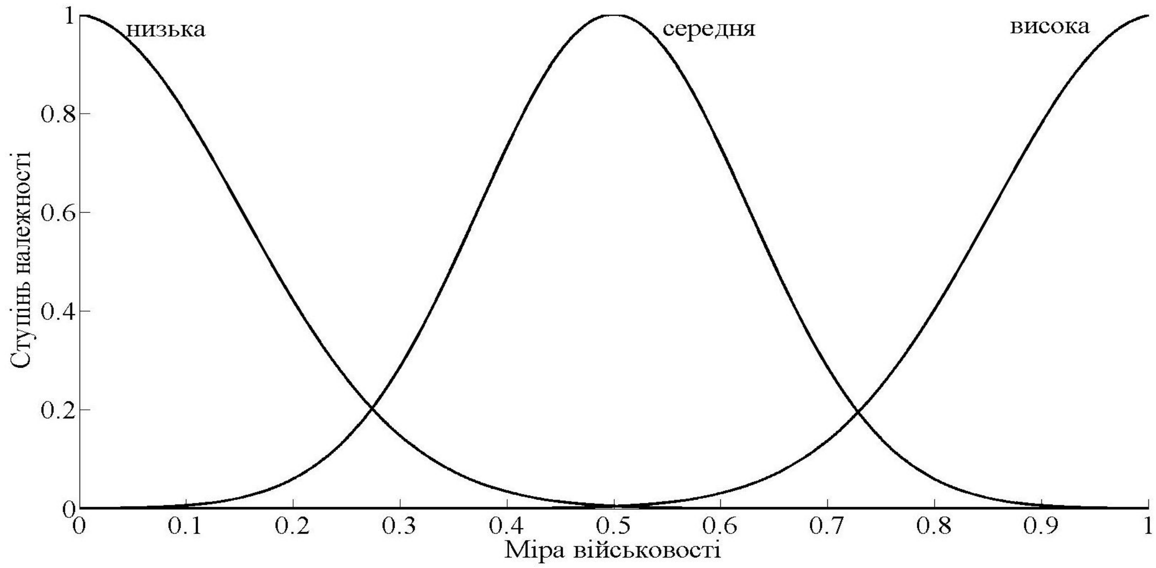


Рисунок 3 – Функції належності для лінгвістичної змінної «Міра військовості»

Для кожного респондента було підраховано його ступінь належності групі військових осіб. Оскільки наявність екстремумів у цільовій поверхні має особливе значення для респондентів із достатньо великою мірою належності, то цільову поверхню було збудовано для наступних інтервалів (рис. 4):

$$\Delta_{\tilde{\mathbf{M}}s} = \{(0,4;0,5],(0,5;0,6],(0,6;0,7],(0,7;0,8],(0,8;0,9],(0,9;1,0]\}.$$

Отримана цільова поверхня має локальні екстремуми, зокрема, в таких точках, як (06700, (0,9;1,0]), (06700, (0,4;0,5]), (06200, (0,8;0,9]) тощо. Розв'язок задачі групової анонімності можна отримати шляхом перенесення екстремумів з одного параметризуючого значення в інше або створення додаткових, на фоні яких початкові буде замасковано. При цьому потрібно зберегти корисність даних. У даній роботі було використано апарат вейвлет-перетворень, який дозволяє розв'язати задачу групової анонімності і при цьому зберегти високочастотні особливості цільової поверхні [10]. Застосовуючи койфлет 5 порядку, можна отримати модифіковану цільову поверхню (рис. 5).

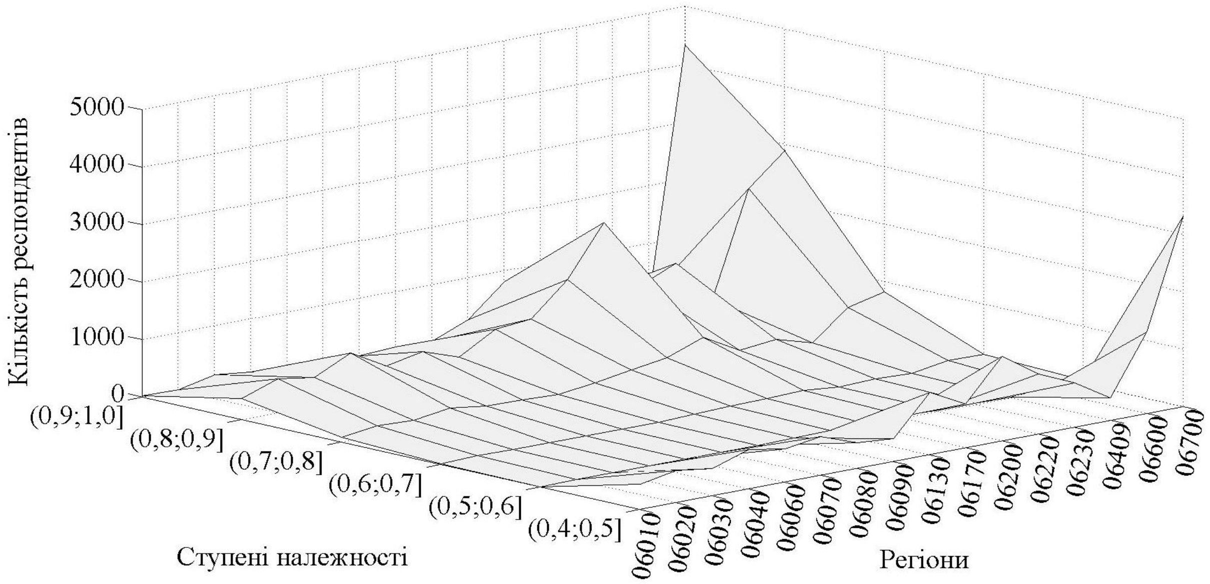


Рисунок 4 – Цільова поверхня для прикладу

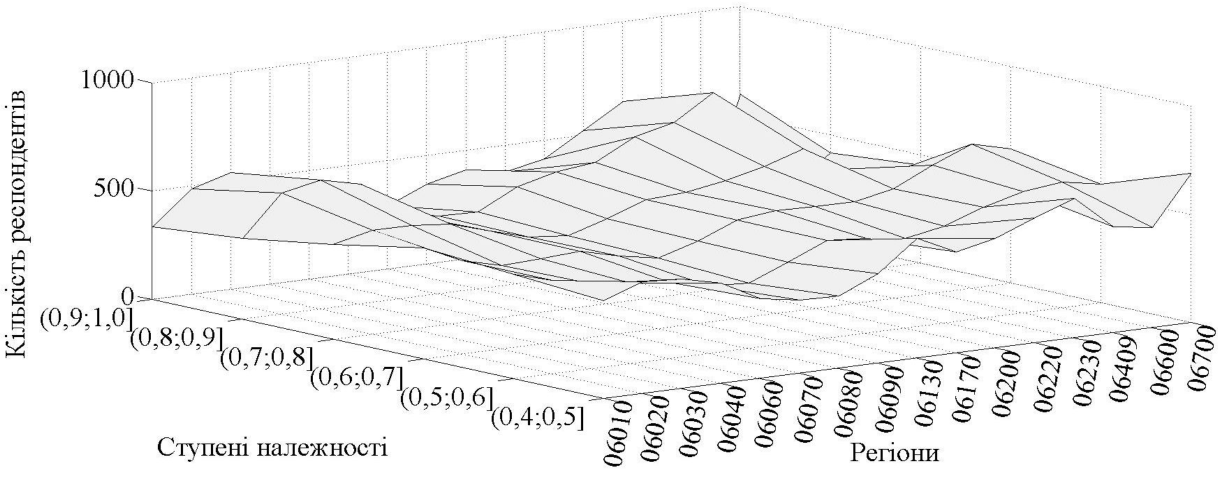


Рисунок 5 – Модифікована цільова поверхня для прикладу

Даний розв'язок задачі групової анонімності є прийнятним, оскільки всі максимуми або перенесено в інші точки на поверхні, або замасковано новими екстремальними значеннями.

Висновки

У даній роботі було запропоновано схему забезпечення нечіткої групової анонімності даних перепису населення з урахуванням їхньої лінгвістичної природи. Детально описано процедуру врахування таких даних для побудови поверхні, яка показує статистичний розподіл респондентів з групи ризику.

Перспективними, на нашу думку, є дослідження в напрямку покращення методів модифікації цільової поверхні з метою мінімізації пошкодження якості (корисності) вихідного мікрофайла, а також застосування нейромережних технологій до автоматизації підбору функцій належності для атрибутів мікрофайла.

Література

1. IPUMS: Minnesota Population Center. Integrated Public Use Microdata Series International [Електронний ресурс]. – Режим доступу : https://international.ipums.org/international/.
2. A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.34 [Електронний ресурс] / A. Pfitzmann, M. Hansen. – 2009. – Режим доступу : http://dud.inf.tu– dresden.de/Anon_Terminology.shtml.
3. Zadeh L.A. Outline of a New Approach to the Analysis of Complex Systems and Decision Processes / L.A. Zadeh // IEEE Transactions on Systems, Man, and Cybernetics. – 1973. – Vol. SMC–3, № 1. – P. 28-44.
4. Chertov O. Statistical Disclosure Control Methods for Microdata / O. Chertov, A. Pilipyuk // Intern. Symposium on Computing, Communication and Control. – Singapore : IACSIT, 2009. – P. 338-342.
5. Privacy – Preserving Data Publishing: A Survey on Recent Developments / B. Fung, K. Wang, R. Chen, P. Yu // ACM Computing Surveys. – 2010. – Vol. 42(4). – P. 1-53.
6. Dwork S. Differential Privacy / S. Dwork // Automata, Languages and Programming [ed. M. Bugliesi, B. Preneel, V. Sassone, I. Wegener]. – Berlin ; Heidelberg : Springer, 2006. – LNCS, vol. 4052. – P. 1-12.
7. Chertov O. Group Anonymity / O. Chertov, D. Tavrov // IPMU – 2010 ; [ed. E. Huellermeier., R. Kruse]. – Heidelberg : Springer, 2010. – CCSI, vol. 81. – P. 592-601.
8. Чертов О.Р. Group Anonymity: Problems and Solutions / О.Р. Чертов, Д.Ю. Тавров // Інформаційні системи та мережі. – Львів : Вид-во Львівської Політехніки, 2010. – № 673. – С. 3-15.
9. Чертов О.Р. Providing Data Group Anonymity Using Concentration Differences / О.Р. Чертов, Д.Ю. Тавров // Математичні машини і системи. – 2010. – № 3. – С. 34-44.
10. Chertov O. Group Methods of Data Processing / Chertov O. – Raleigh : Lulu.com, 2010. – 156 p.
11. Chertov O. Data Group Anonymity: General Approach / O. Chertov, D. Tavrov // International Journal of Computer Science and Information Security. – 2010. – Vol. 8(7). – P. 1-8.
12. Zadeh L.A. Fuzzy Sets / L.A. Zadeh // Information and Control. – 1965. – № 8. – P. 338-353.
13. Zadeh L.A. The Concept of a Linguistic Variable and its Application to Approximate Reasoning / L.A. Zadeh // Information Sciences. – 1975. – № 8. – P. 199-249.
14. Zadeh L.A. Fuzzy Logic, Neural Networks, and Soft Computing / L.A. Zadeh // Communications of the ACM. – 1994. – Vol. 37, № 3. – P. 77-84.
15. Klir G.J. Fuzzy Sets and Fuzzy Logic: Theory and Applications / G.J. Klir, B. Yuan. – Prentice Hall, 1995. – 592 p.
16. Cano I. Evaluation of Information Loss for Privacy Preserving Data Mining Through Comparison of Fuzzy Partitions / I. Cano, S. Ladra, V. Torra // FUZZIEEE2010, IEEE, Barcelona. – 2010. – P. 1-8.
17. Mamdani E.H. An experiment in linguistic synthesis with a fuzzy logic controller / E.H. Mamdani, S. Assilian // International Journal of Man – Machine Studies. – 1975. – Vol. 7, № 1. – P. 1-13.
18. Круглов В.В. Нечеткая логика и искусственные нейронные сети / В.В. Круглов, М.И. Дли, Р.Ю. Голунов. – М. : Физматлит, 2001. – 221 с.
19. U.S. Census 2000. 5 – Percent Public Use Microdata Sample Files [Електронний ресурс]. – Режим доступу : http://www.census.gov/census2000/PUMS5.html.

Literatura

1. IPUMS: Minnesota Population Center. Integrated Public Use Microdata Series International https://international.ipums.org/international/.
2. A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.34. 2009. http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.
3. Zadeh L. IEEE Transactions on Systems, Man, and Cybernetics. 1973. Vol. SMC-3, No. 1. P. 28-44.
4. Chertov O. Intern. Symposium on Computing, Communication and Control. Singapore: IACSIT. 2009. P. 338-342.
5. Fung B. ACM Computing Surveys. 2010. Vol. 42(4). P. 1-53.
6. Dwork S. Automata, Languages and Programming. Berlin/Heidelberg: Springer. 2006. LNCS. Vol. 4052. P. 1-12.

7. Chertov O. IPMU-2010. Heidelberg: Springer. 2010. CCSI. Vol. 81. P. 592–601.
8. Chertov O. R. Informacijni systemy ta merezhi. L'viv : Vid-vo L'vivs'koi Politehnyky. 2010. № 673. S. 3-15.
9. Chertov O. R. Matematychni mashyny i systemy. 2010. № 3. S. 34-44.
10. Chertov O. Group Methods of Data Processing. Raleigh: Lulu.com. 2010. 156 p.
11. Chertov O. International Journal of Computer Science and Information Security. 2010. Vol. 8(7). P. 1-8.
12. Zadeh L. A. Information and Control. 1965. 8. P. 338-353.
13. Zadeh L. A. Information Sciences. 1975. 8. P. 199–249.
14. Zadeh L. A. Communications of the ACM. 1994. Vol. 37. №. 3. P. 77-84.
15. Klir G. J. Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall. 1995. 592 p.
16. Cano I. FUZZIEEE2010, IEEE. Barcelona. 2010. P. 1-8.
17. Mamdani E. H. International Journal of Man-Machine Studies. 1975. Vol. 7. №1. P. 1-13.
18. Kruglov V. V. Nechetkaja logika i iskusstvennye nejronnye seti. M.: Fizmatlit. 2001. 221 s.
19. U.S. Census 2000. 5-Percent Public Use Microdata Sample Files. http://www.census.gov/census2000/PUMS5.html.

RESUME

O.R. Chertov, D.Y. Tavrov

Providing Group Anonymity in the Microfile with Fuzzy Data

Contemporary information technologies make it possible to analyze large amounts of digital data, especially primary non-aggregated data, which can be used to generate hypotheses and validate them. The IPUMS-International project, which contains about 397 million records about respondents of 185 censuses in 62 countries, is a prominent example of accessibility of the primary data. But, publishing these data obviously increases the threat of disclosing sensitive information.

Protecting information on a single person implies providing *individual* data anonymity. Techniques for carrying it out are widely known as the part of privacy-preserving data mining, statistical disclosure control, and other scientific fields. Protecting intrinsic data properties and distributions is achieved by providing data *group* anonymity. Group anonymity methods aim at protecting those data features which are not distinguishable by only analyzing standalone data elements.

All the existing methods for providing data anonymity process numeric data which do not always adequately describe underlying data. Although, it is known that the key elements of the human conscience are fuzzy sets, i.e. classes of objects with a gradual grade of membership rather an abrupt one. Therefore, many results of statistical researches are essentially linguistic. Thus, it is convenient to apply fuzzy logic techniques to providing data group anonymity.

In this paper, we propose the method for utilizing fuzzy information in census microfile data for providing their group anonymity. We also illustrate this approach with real-life example.

Стаття надійшла до редакції 01.06.2012.